Modelling speciation-fossilization process in total-evidence dating and its application to penguin evolution

Alexandra Gavryushkina
T. Heath, D. Ksepka, T. Stadler, D. Welch, and A. Drummond

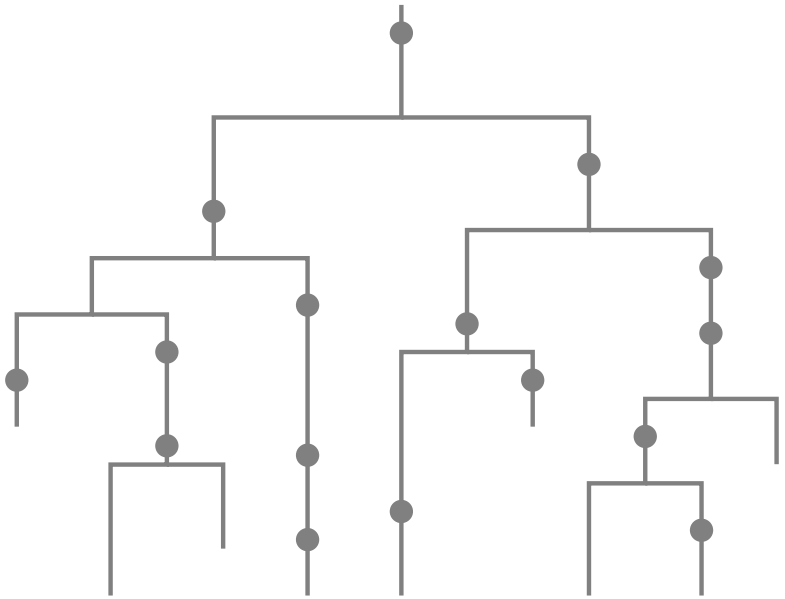
The University of Auckland
ETH Zürich

GSA 2016

In this presentation

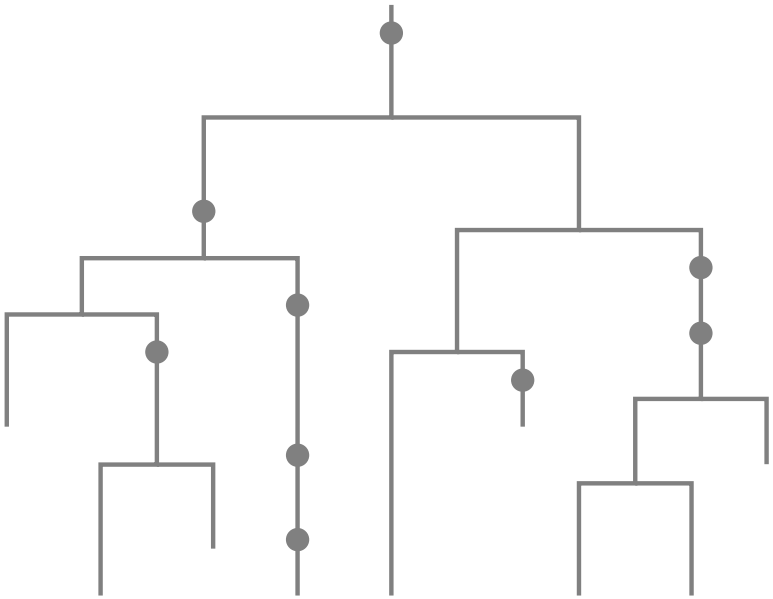
- A statistical Bayesian method to infer dated phylogenies known as total-evidence or 'tip-dating' method.
- Modelling speciation-fossilization process.
- Application of the method to a penguin dataset.

Evolutionary process



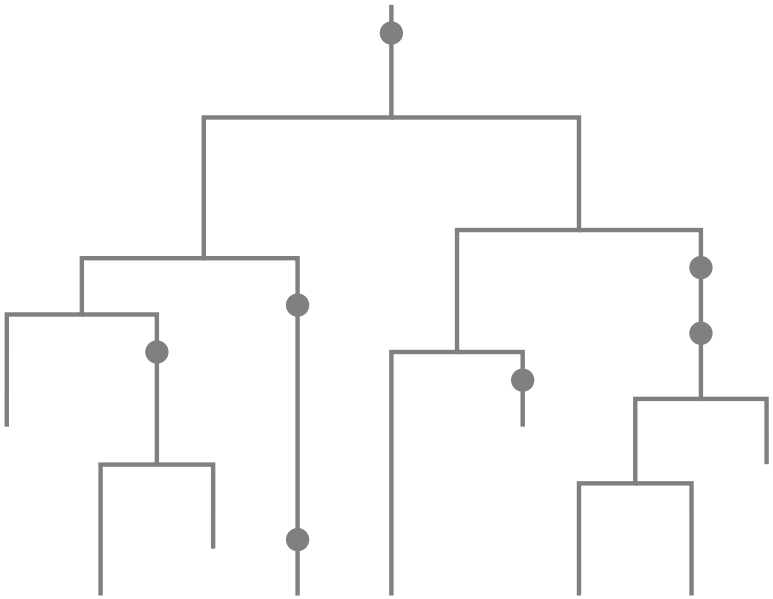
Full tree

Evolutionary process



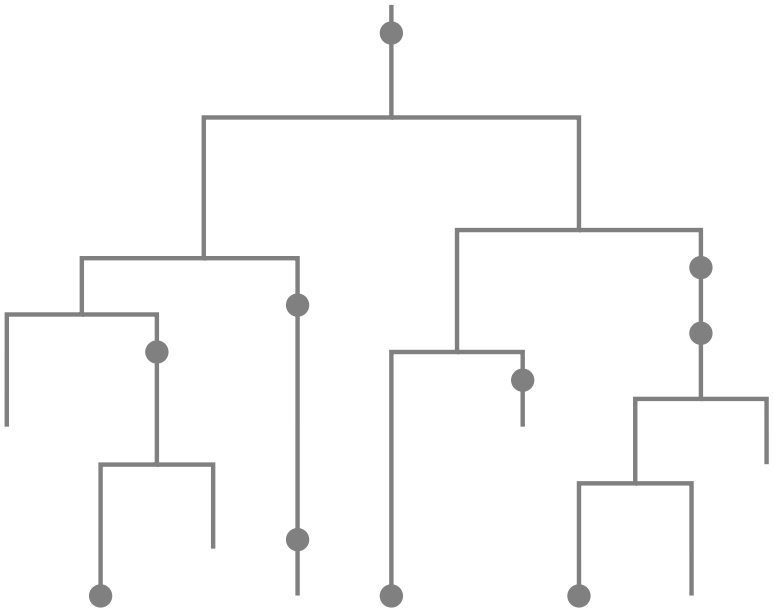
Full tree

Evolutionary process



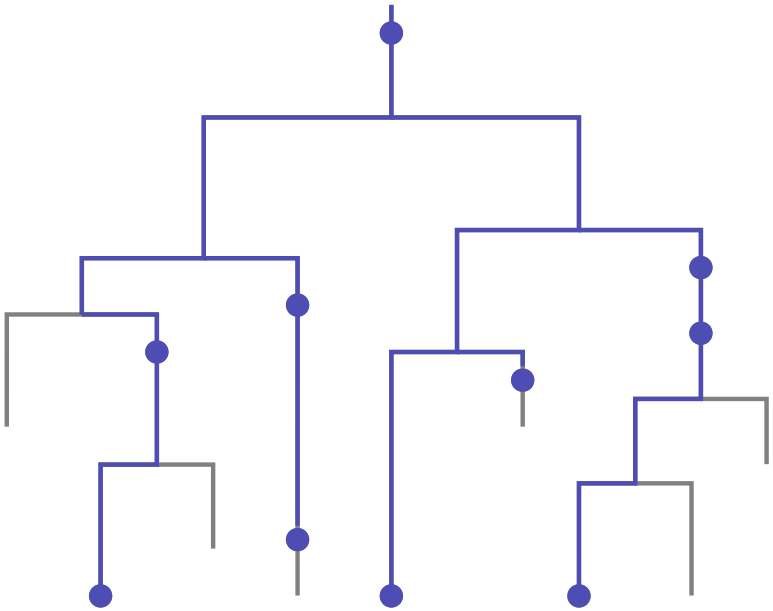
Full tree

Evolutionary process



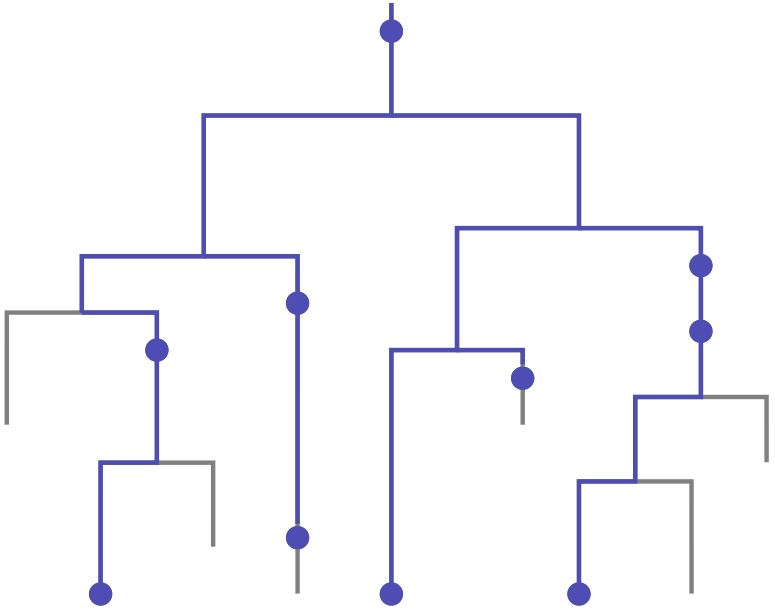
Full tree

Evolutionary process

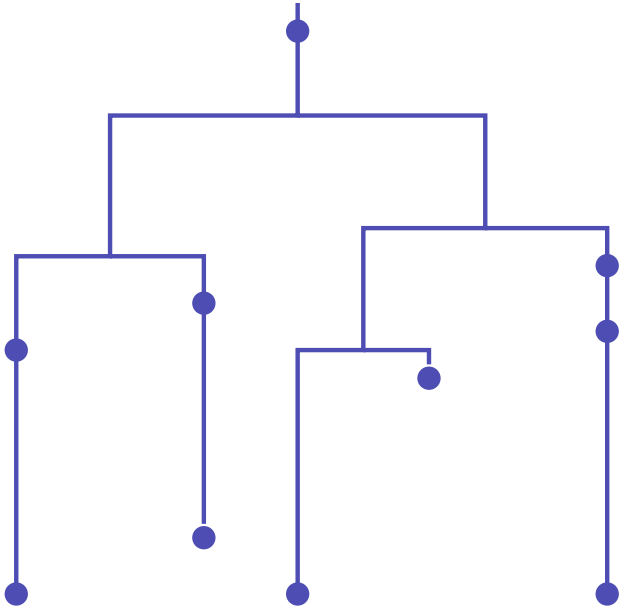


Full tree

Evolutionary process



Full tree



Sampled tree

Data

The data we observe are:

 molecular sequences

 morphological matrix (fossil and extant species)

 fossilization dates

- We use the models that describe these processes to infer dated phylogenies from these data.
- We co-estimate topologies and divergence dates.

Bayesian approach

D	morphological and molecular data,
$\bar{\tau}$	fossil occurrence intervals (either reflecting the uncertainty in the age estimate or representing the fossil taxon age range),
$\mathcal{T}$	phylogeny (topology with node ages),
$\bar{\eta}$	tree model parameters, and
$\bar{\theta}$	substitution and clock model parameters.

Using MCMC we sample from the posterior distribution:

$$f(\mathcal{T}, \bar{\theta}, \bar{\eta} | D, \bar{\tau}) \propto f(D|\mathcal{T}, \bar{\theta}) f(\bar{\tau}|\mathcal{T}) f(\mathcal{T}|\bar{\eta}) f(\bar{\eta}) f(\bar{\theta}) \propto$$
$$f(D|\mathcal{T}, \bar{\theta}) \delta(\mathcal{T} \in T_{\bar{\tau}}) f(\mathcal{T}|\bar{\eta}) f(\bar{\eta}) f(\bar{\theta}),$$

where $T_{\bar{\tau}}$ is the set of phylogenies that are consistent with intervals $\bar{\tau}$ and we assume that

$$f(\bar{\tau}|\mathcal{T}) \propto \delta(\mathcal{T} \in T_{\bar{\tau}})$$

Terminology

Joint inference — **joint analysis** of

- comparative data (morphological and/or molecular) and
- temporal data (fossil occurrence dates)

co-estimating topology and divergence dates

When both molecular and morphological data are used in a joint inference it is called **total-evidence**.

Challenges of the method

The first attempts to apply the method produced very old divergence date estimates and the method was much criticised.

There are two main direction for improving the method:

- Improving the modelling of the morphological evolution because the models that are currently used were initially developed for molecular evolution.
- Improving the modelling of the speciation-fossilisation process. The choice of the model generating the tree is very important because unlike the molecular sequences, morphological data of fossils are limited and the assumptions of the model strongly influence the results.

Speciation-fossilization models

Only a few models have been implemented for the joint inference to date. Most of the models are variants of the birth-death model with or without sampling.

1. Yule model (pure birth without sampling)
2. Uniform model (not a birth-death model)
3. Birth-death model (no sampling)
4. Birth-death-sampling model (fossilized birth-death model, FBD)
5. Skyline FBD
6. Diversified skyline FBD

Fossillized birth-death model (FBD)

Stadler 2010, Heath et al 2014.

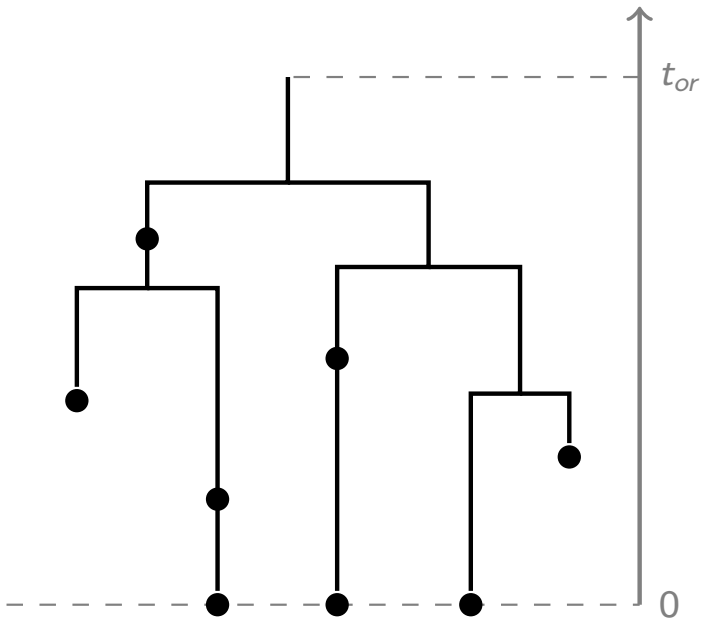
The process starts at time $t_{or} > 0$ and ends at time zero (present time).

- birth rate λ
- death rate μ
- sampling rate ψ
- sampling at present probability ρ

Model parameters: $\eta = (t_{or}, \lambda, \mu, \psi, \rho)$.

All the parameters are identifiable.

No constraints on the root or origin age are required.



Sampled tree

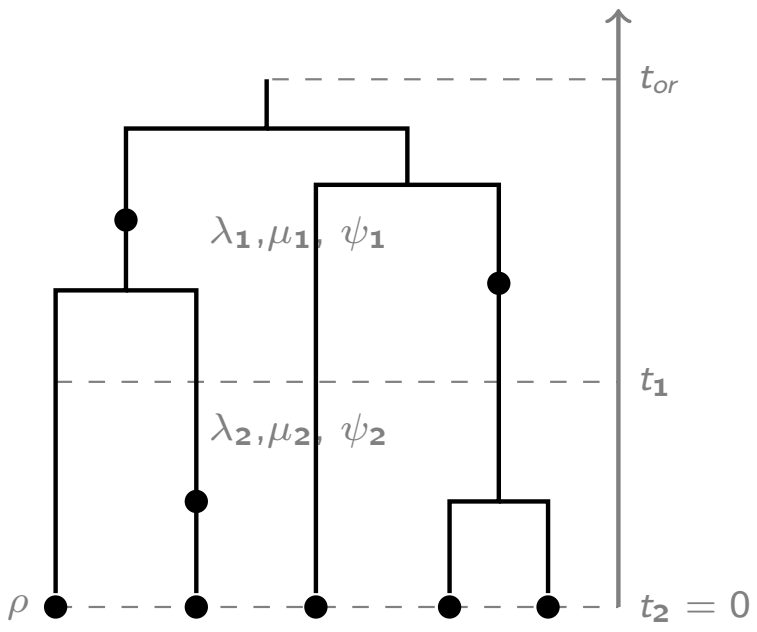
Skyline FBD

Stadler *et al.* (2012),
Gavryushkina *et al.* (2014)

There are k time intervals and parameters remain constants within the intervals but may vary from one interval to another

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_k$
- sampling at present ρ

Model parameters: $\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho)$



Sampled tree

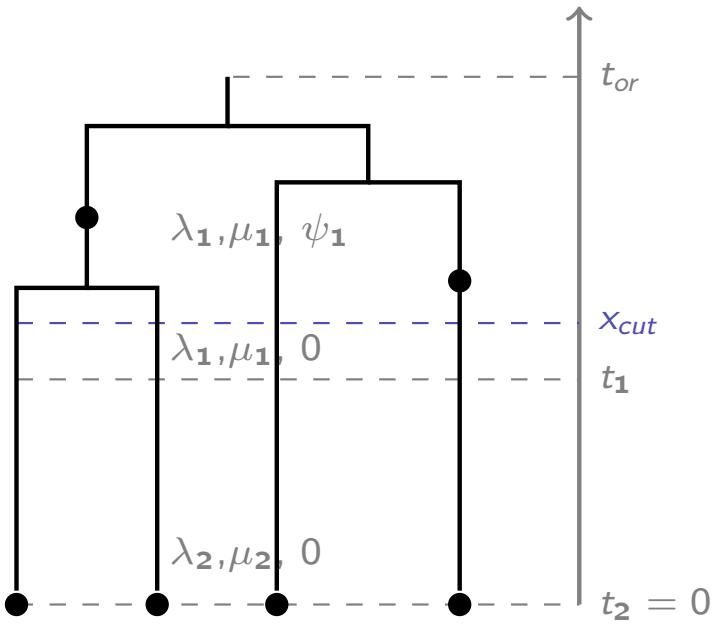
Diversified skyline FBD

Hönna *et al.* (2011) and Zhang *et al.* (2016)

There is a cut-off time x_{cut}. There are no fossil samples after x_{cut} and a single descendant (if any) of every branch existing at time x_{cut} is sampled at present.

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_m, 0, \ldots, 0$

Model parameters:
$\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \psi_1, \ldots, \psi_m)$



Sampled tree

Influence of the speciation-fossilization model

Matzke and Wright (2016) analysis of fossil Canidae:

	Canidae	crown Caninae	crown *Canis*
Uniform	49 Ma	38.9 Ma	27.5 Ma
FBD	36.3 Ma	9.8 Ma	2.8 Ma

Zhang *et al.* (2016) analysis of Hymenoptera + outgroups:

	Hymenoptera
Uniform	306 Ma
Skyline FBD	346.6 Ma
Diversified Skyline FBD	251.7 Ma

Penguin analysis

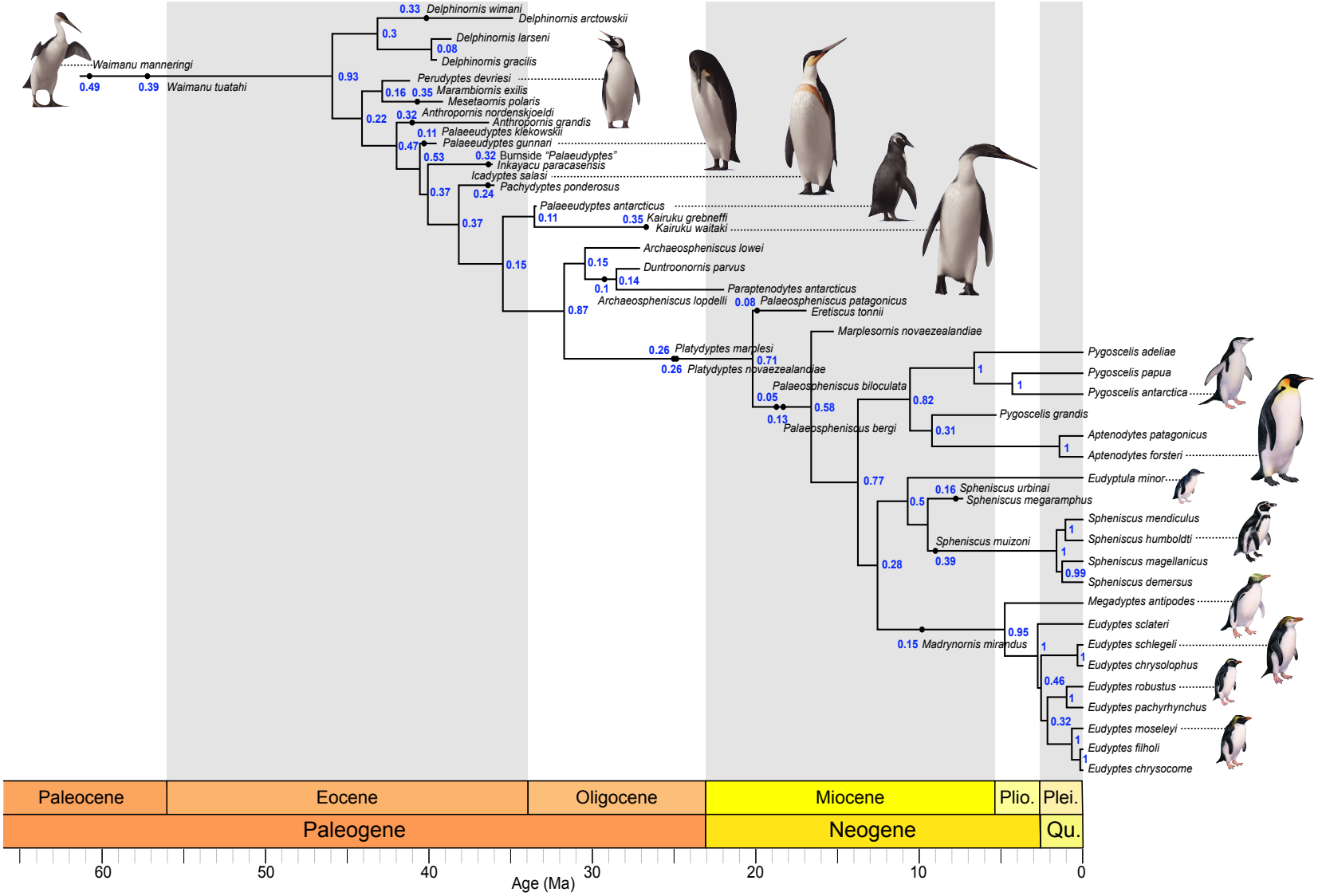
We applied this method to analyze a penguin dataset from Ksepka *et al.* (2011):

- morphological matrix of 36 fossil 19 extant species,
- molecular data of extant species, and
- fossil occurrence intervals

We used

- different variants of Lewis Mk model for morphological evolution,
- two independent clock models (relaxed or strict) for molecular and morphological data, and
- FBD model with uninformative prior distributions for the parameters with ρ fixed to one.

Maximum sampled ancestor clade credibility tree of penguins



Estimates of the penguin crown age

Baker *et al.* (2006):
40.5 Ma, CI: [34.2,47.6]

Brown *et al.* (2008): 50 Ma

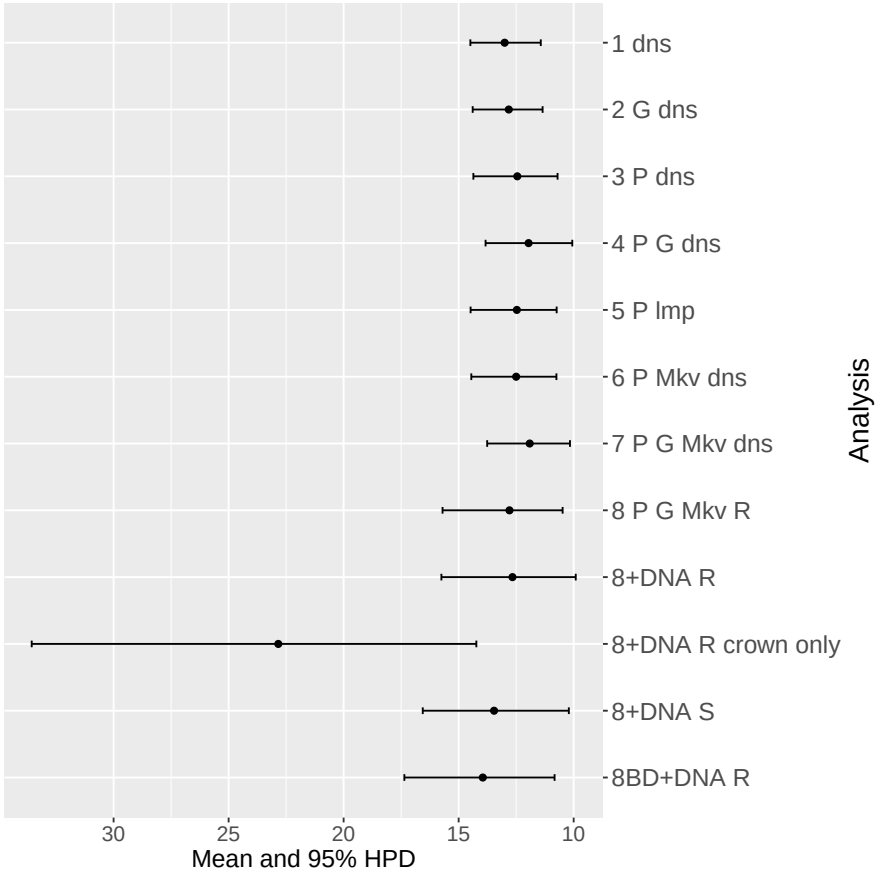
Subramanian *et al.* (2013):
20.4 Ma, HPD: [17,23.8]

Jarvis *et al.* (2014) and
Li *et al.* (2014):
23 Ma, CI: [6.9,42.8]

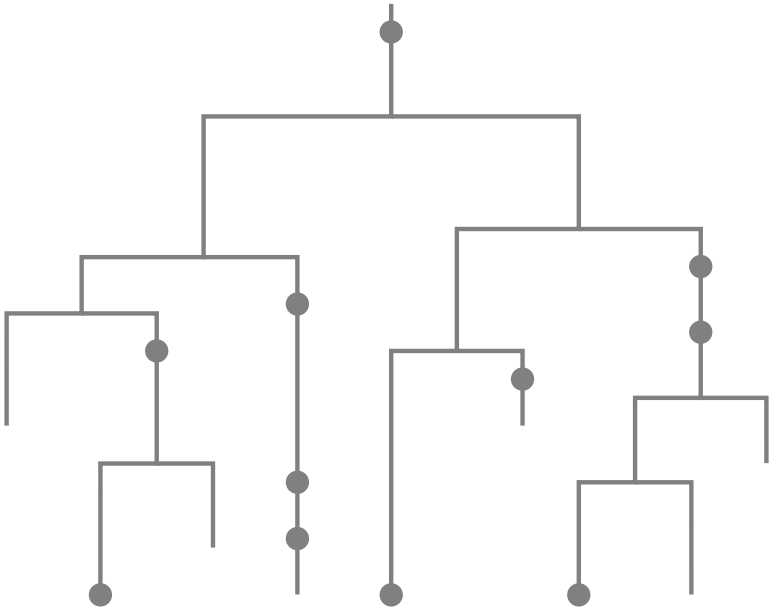
Our estimate:

12.7 Ma, HPD: [9.9, 15.7]

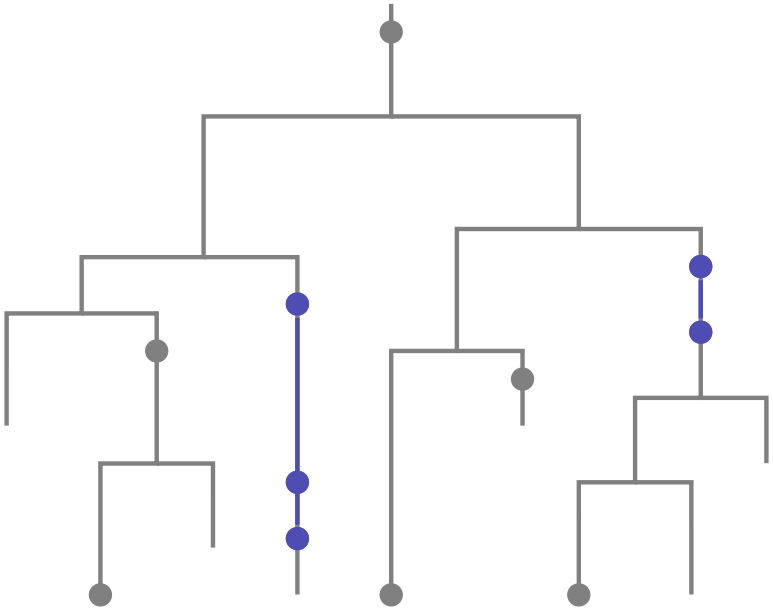
Our estimate without stem fossils:
22.8 Ma, HPD: [14.2, 33.6]



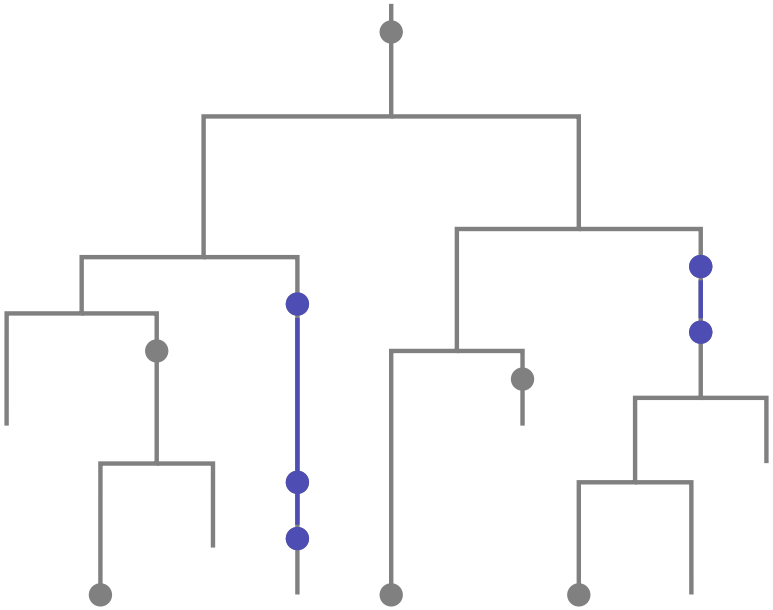
Improved modelling of fossil sampling process



Improved modelling of fossil sampling process



Improved modelling of fossil sampling process



Incorrect modeling:

- we replaced several fossil samples of the same taxon with just one and assumed its age ranges between the first and the last occurrences.

Improved modeling:

- we include all occurrences as input data or
- we only include the first and the last occurrences and modify the model accordingly.

Summary

- The amount and quality of fossil occurrence data and the models that describe fossilization process greatly influence estimated phylogenies.
- The models that do not describe fossil sampling process are not recommended.
- The variants of FBD model are useful and should be used appropriately.
- FBD model is sensitive to biased sampling. Thus, we should account for diversified sampling. More accurate modelling of fossil sampling (e.i., accounting for multiple samples of the same taxon) might improve the inference.
- Including more fossils, e.i., stem fossils, can greatly improve the results.

Software

The method is available in BEAST2 (`beast2.org`)
with packages:

- SA (enables sampled ancestor trees and FBD model)
- MM (adds models of morphological evolution)
- BDSKY (adds FBD skyline model)